Service Science and Management Research (SSMR) Volume 4, 2015 www.seipub.org/ssmr
doi: 10.14355/ssmr.2015.04.008

An Effective and Efficient Feature Selection Scheme

Diao Xingchun[1], Zhou Xing[2], Qiu Hangping[3], Cao Jianjun[4]

PLA University of Science and Technology

No.18, Houbiaoying Road, Qinhuai Qu, Nanjing city, Jiangsu Province, China

[1]diaoxch640222@163.com; [2] zx0327@163.com; [3]qiuhp8887@hotmail.com; jianjuncao@yeah.net

Abstract

Affinity Propagation (AP) is widely used since being put forward, this paper shows advantages of using it in duplicate detection. To further enhance the effectiveness and efficiency of AP, a hybrid feature selection scheme is proposed, that is, a filter feature subset evaluation criterion is chosen to decide the best subsets for a given cardinality, thus reducing the complexity, then using the performance of AP as a criterion for feature subset evaluation criterion to select the final best subset among the best subsets across. In case of data with large number of features, sampling search and simulated annealing are used to reduce the time consuming. Experiments show that this scheme works with both effectiveness and efficiency. The sample search can achieve a good effectiveness with a relatively good efficiency, and simulated annealing can achieve a not bad effectiveness with good efficiency.

Key words:

Affinity Propagation; Feature Selection; Hybrid; Duplicate Detection

Introduction

Duplication detection is a long lasting problem in data cleaning since first being put forward in 1959 as record linkage [1], many classification algorithms have been used to distinguish whether a record group is duplicate or not, including Bayes [1], rule based classification [2], decision tree [3], SVM [4], etc. While classification algorithms are used to tell apart duplicate records and distinct records, another intuition is to use clustering to group together similar records, which can work as a complete duplicate detection process or as a preprocessing for classification like the blocking in [5]. Thus, clustering for duplicate detection is worth of attentions. Oktie Hassanzadeh showed the use of unconstrained clustering to group potential duplicates, and provided a framework to evaluate the quality of the resulting clusters [6], the unconstrained clustering means that the number of clusters is not needed to be specified in advance, which fits for most duplicate detection applications for the fact that the number of duplicate record groups is not known previously.

Affinity Propagation (AP) is a newly proposed unconstrained clustering algorithm and it takes the similarity of pair wise instances as input [7], which enlarges its input, meaning that the inputting type of records and fields to be compared is not needed to be adjusted for the particular algorithm, just the similarity between a record pair needed, making AP quite fit for the duplicate detection, since in most cases, the similarity calculation is a basic step. Besides it is also an unconstrained clustering algorithm and not sensitive to the initialized points, and it has a relatively higher accuracy than other clustering algorithms, the above peculiarities making it suit for duplicate records detection.

The algorithm stops under three conditions: after a fixed number of iterations, after the message change falls below a threshold or after the decision stay constant.

However, in case that each instance has many attributes, the efficiency decays, and with the presence of duplications, the effectiveness of clustering algorithms also decreases. A natural way to amend this situation is to use feature selection since it can filter out redundant and irrelevant features, so as to improve effectiveness and efficiency.

Feature selection is the process of selecting a subset of original features, with commonly four procedures: feature

subset generation, feature subset evaluation, stop criteria, and result validation [8], many researches on feature selection focus on subset evaluation and subset search like [9][10][11]. This work also focuses on subset evaluation and subset search.

Feature subset evaluation can be divided into three categories, i.e., filter, wrapper, and hybrid based on whether a particular learning algorithm is involved. Filter methods do not depend on the particular learning algorithm and they use the general characteristics of data, while wrapper methods do, they use the performance of a predefined learning algorithm as evaluation criterion, but its efficiency also falls down due to the algorithm involved and may be not suitable for other algorithms. Hybrid methods try to take advantages of both methods by using different evaluation criteria in different stages, often they can be expressed as: in each subset of cardinality c, they use filter method to get the best subset with c+1 cardinality and use wrapper method to compare the effectiveness of the current best subset of c+1 with the previous best subset of cardinality c so as to get the overall best subset.

The feature subset search has three categories too, as: complete search, sequential search and random search. Complete search can guarantee finding the best subset, but it has a high complexity as $o(2^N)$. Sequential search means giving up completeness and searching in sequential order, like adding or removing features one by one or add p features while remove q features, and its complexity is often $o(N^2)$. The random search means starting at a random point and then generates next subset in sequential order or random order. More detailed information about feature subset evaluation and subset search can be found in [8].

To improve the performance of AP, the most intuitive way is to use wrapper method, that is to use the performance of AP algorithm as a subset evaluation criterion to conduct feature selection for the property of wrapper method, but it has the disadvantage of computation complexity. At present, no wrapper work has been done to improve AP clustering as we know, while filter methods cannot guarantee the features being selected the best suited for AP algorithm. In situations where the effectiveness of the particular algorithm is more preferred, the filter method will not work that well for a particular algorithm.

Hybrid methods will thus be the most suited for this situation, which combine the advantages of filter and wrapper methods.

This work tries to use hybrid method to do feature subset. That is to use filter methods to get the best subset of cardinality c, and wrapper methods to get the best overall subset. As to wrapper method, the performance of AP is evaluated using the F score in [6] in case of ground truth present. Moreover in subset cardinality changes, sequential search and random search are used as subset search.

Currently, there are many filter evaluation criteria, varying in fitness and generalization. In this paper, several filter criteria are compared and the best one is chosen to rank features so that the ranked features can help find the best feature subset within a relatively small time and fit for as much data type as possible. What's more, the best overall subset is not easily obtained due to multi-extreme values present, which will be shown in later parts using experiments and is solved using sample search and simulated annealing.

Using four benchmark datasets including: two classification datasets from UCI (breast cancer [12]and dermatology [13]) and another two datasets formally used to evaluate duplicate detection algorithm in [6](warpPIE10P and pixraw10P), we showed that for datasets with duplicate records, the very method can have a better clustering performance than those not, the efficiency would also increase due to feature reduction. The sample search can always have a good effectiveness, while the simulated annealing sometimes can achieve a good and quick result and sometimes a not bad result due to parameters sensitiveness. For datasets without duplicate records, the effectiveness can also get improved, but not as much as that with duplications.

The following is feature subset evaluation criterion chosen in part 2, search procedure in part 4, overall comparison in part 4, and conclusion in part 5.

Feature Subset Evaluation Criteria

Since feature ranking is expected using a filter criterion, in this part, first, several typical evaluation criteria are

referred to, then the best is chosen according to performance comparison of three typical score functions using benchmark data.

Related Work

There are currently many filter feature subset evaluation criteria, including variance, laplacian score [14], feature similarity [15], constraint score [16], distance[8], fisher score [17], relevant and redundant [11], information gain[8] and dependency measure [8].

Among the criteria mentioned above, variance, laplacian score and fisher score are three typical score functions and usually used in feature selection [16], they can be used to rank features. In this paper, they are compared using benchmark data for fitness and generalization so that the best is chosen for ranking the features. Furthermore, in computing laplacian score, both situations with class label present and without label are considered in constructing the neighborhood graph.

Theoretically, laplacian score would be more suited than variance, because of its locality preserving ability, besides its computation does not need clustering center involved, making it more robust.

The definitions of laplacian score and fisher score are given as follows:

The Laplacian score of the *r*-th feature is computed as follows:

$$L_r = \frac{\sum_{i,j}(f_{ri} - f_{rj})^2 S_{ij}}{\sum_i (f_{ri} - f_{rj})^2 D_{ii}}$$

Where, f_r is the *r*-th feature, f_{ri} and f_{rj} are the *i*-th and *j*-th sample of f_r respectively. *D* is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$, and S_{ij} is defined by the neighborhood relationship between samples.

The S_{ij} has several definitions, among a simplest form is as follows [9]:

$$S_{ij} = \begin{cases} 1, & x_i \, is \, a \, neighbor \, of \, x_{j,} \, or \, x_j \, is \, a \, neighbor \, of \, x_i \\ 0, & otherwise \end{cases}$$

Another form can be found in [14]

Fisher score is computed as follows:

$$F_r = \frac{\sum_{i=1}^{c} n_i (\mu_r^i - \mu_r)^2}{\sum_{i=1}^{c} n_i (\sigma_r^i)^2}$$

Where μ_r^i and σ_r^i are the mean value and variance of the cluster *i* samples of the *r*-th feature respectively, *i* varies from 1 to *c*. n_i is the number of samples of feature *r*, and μ_r the mean value of all samples of feature *r*.

Criteria Comparison

To further compare the three criteria, four experiments are conducted using the datasets as mentioned previously.

For the breast cancer dataset, since its first attribute is id, whose similarity does not help distinguish instances, thus is discarded. Besides, those instances with missing attributes are also discarded for breast cancer and dermatology datasets.

Since the two datasets from Toronto university have many features, they are sampled, for warpPIE10P data, it has 2420 features and is sampled every 20 features, the pixraw10P has 10000 features and is sampled every 100 features.

For laplacian score without class labels, AP algorithm is first conducted to generate clusters such that those within the same cluster are regarded as neighbors, in case of class label present, those with the same class label are regarded as neighborhoods.

After ranking the features using the particular criterion, the feature subsets then add features in forward and backward order, where backward order means the feature subset increases from big to small, the forward from small to big, then AP is conducted and the F is calculated. Followings are the result comparison.

In each figure, the 'plus' refers to fisher score in forward sequential order, the 'x-mark' refers to fisher score in backward sequential order; the 'star' refers to laplacian score with class label in forward order, the 'rectangle' refers to laplacian score with class label in backward order, the 'triangle' refers to laplacian score without class label in forward order, the 'circle' refers to laplacian score without class label in backward order; the 'diamond' refers to variance in forward order, the 'pentagram' refers to variance in backward order, and the 'dash' refers to base score, i.e., the F score obtained without feature selection.

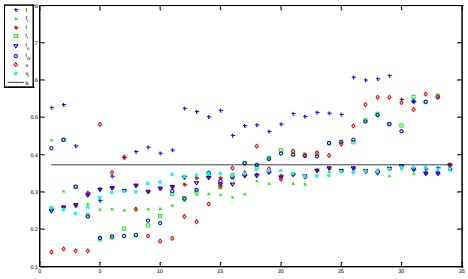

FIGURE 1 COMPARISON OF THE ABOVE CRITERIA USING DERMATOLOGY DATA

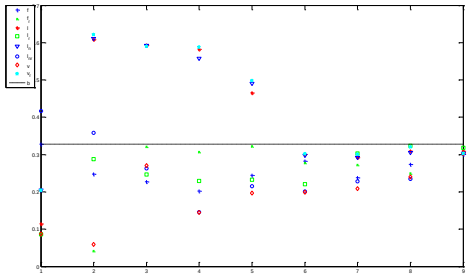

FIGURE 2 COMPARISON OF THE ABOVE CRITERIA USING BREAST DATA

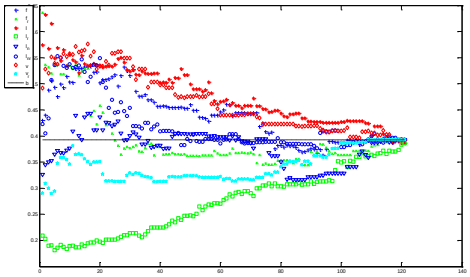

FIGURE 3 COMPARISON OF THE ABOVE CRITERIA USING WARPPIE10P DATA

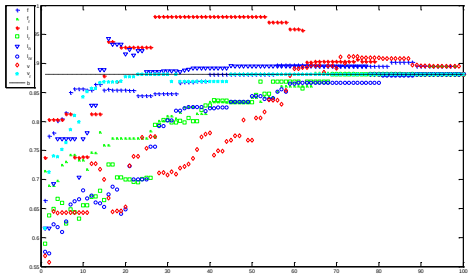

FIGURE 4 COMPARISON OF THE ABOVE CRITERIA USING PIXRAW10P DATA

From the above four figures, we can conclude that laplacian score with class label is the best, since it can achieve a better result for almost all the datasets, in case of no label information, the laplacian score in reverse order is preferred. Another interesting conclusion is that there is not only one extreme value, leading to local extreme values, troubling the subset search procedure.

Feature Subset Search

From the figures above, we can find a big problem for subset searching is that the F does not have only one extreme value, thus leading to local extreme values. Another challenge is that when the feature number is small, it is easy to search along all features, the complexity is still small, just proportional to the number of features. However, when the feature number reaches hundreds even thousand, like the warpPIE10P and pixraw10P data, though still linear, would be time-consuming due to AP algorithm. To solve this problem, two algorithms are introduced for subset search. One is sample search, that is, first search in a coarse granularity, after got the best result, then search in a finer granularity based on an assumption that the F will not have a sudden change due to variation on number of features; the other is to use the simulated annealing [18] to do random search to reduce complexity and avoid local extreme value. The sample search searches along in sequential order, while simulated annealing searches in random order.

The sample interval of sample search varies with the feature number as a tradeoff of effectiveness and efficiency, a

loose factor is used to broaden the search range to further consider accuracy and if the feature number due to loose factor is still big, then it is sampled again. The overall time complexity is O (*kA*), where *k* is the number of features and *A* is the complexity of AP, which can be found in [7]. Table 2 shows the result of sample search, the pixraw10P dataset has an interval of 100, and is sampled twice, the warpPIE10P is sampled using 20 and 24 as interval for interval comparison. The unit of time is the second.

TABLE 1 THE RESULT OF SAMPLE SEARCH

	Time	Number	F
pixraw10P_100	3253.5	2641	0.9805
warpPIE10P_20	1203.02	40	0.6367
warpPIE10P_24	2228.4	40	0.6367

From table 2, we can conclude that as to warpPIE10P data, the sample search can get a good accuracy invariant of interval. The effectiveness of this method, i.e., F here also gets improved compared than the situation of no feature selection.

For the case of simulated annealing, the parameters chosen are vital. Empirically, the initial value is 0.3 times the max search space length. Experiments show the simulated annealing with proper parameters, would be much quicker with a less good F achieved compared to that of sample search. The temperature is set to be 20 for all cases.

TABLE 2 A COMPARISON OF TYPICAL PARAMETERS

	Tolerance	Random Initial Point	Step Factor	Time	F
warpPIE10P	0.04	Xmax*0.1, Xmax*rand	0.3	399.2	0.4408
	0.03	Xmax*0.1, Xmax*rand	0.3	19.86	0.4296
	0.02	Xmax*0.1, Xmax*rand	0.2	333.2	0.4336
	0.03	Xmax*0.2, Xmax*rand	0.2	28.39	0.4192
	0.03	Xmax * 0.2, Xmax * 0.3	0.2	548.7	0.4036
pixraw10P	0.03	Xmax*0.3,Xmax*rand	0.2	19.6	0.8132
	0.03	Xmax*0.1, Xmax*rand	0.1	697.4	0.9373

From table 3, we can conclude that, the simulated annealing cannot guarantee the best effectiveness achieved especially within a relatively short time, but a long iterative time will lose its advantage comparing to sample search. It is sensitive to parameters and the result is random. Thus in most cases the sample search is more recommended.

Overall effectiveness

To compare the overall effectiveness of the proposed feature selection, a comparison is conducted, where SVM, AP, and AP with feature selection are compared using F score. While using SVM, we divide the dataset into train data and test data, which is the first 600 of breast cancer as train data, the first 300 of dermatology as train data, the first 8 of pixraw10P's each cluster as train data, the first 18 of warpPIE10P 's each cluster as train data. And while conducting experiments, we vary the parameter of RBF kernel from 1 to 100 to get the best result. The overall comparison is as follows:

From the table above, we can see that, the proposed feature selection is always better than AP without feature selection, and in case of duplication present, even better than SVM, thus proving the effectiveness of this scheme. That SVM performs poorly lies in duplication, AP is much better than SVM for duplication dataset because its input is similarity.

TABLE 3 OVERALL EFFECTIVENESS COMPARISON

	SVM	AP	AP with feature selection
breast cancer	0.9495	0.3093	0.6086
dermatology	0.5	0.3616	0.3692
pixraw10P	0.15	0.8805	0.9805
warpPIE10P	0.2	0.3927	0.6367

Conclusion

AP is fit for duplicate detection, since it is an unconstrained clustering algorithm and takes pair wise similarity as input. In this paper, a hybrid feature selection for AP is proposed, it seeks to get a relatively good effectiveness and efficiency at the same time. To achieve this goal, the features are ranked first using a properly chosen score function called laplacian score, which is the filter procedure in hybrid feature selection. Then using sequential search and random search to conduct subset search, which is the wrapper procedure. Considering the case of many features, sample search and simulated annealing are introduced.

Experiment shows, laplacian score is more preferred than others because of its fitness and generalization, the sample search works well for data with large number of features, and can get a good effectiveness, the simulated annealing can sometimes work well with parameters properly chosen. The overall effectiveness using sample search is even better than SVM in case of duplicate present.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China under Grant NO.61371196.

References

[1] Newcombe, H., Kennedy, J., Axford, S., James, A.: "Automatic linkage of vital records", Science ,1959,130(3381), 954–959

[2] Y.R. Wang and S.E. Madnick, "The Inter-Database Instance Identification Problem in Integrating Autonomous Systems", IEEE International Conference on Data Engineering. (ICDE '89),1989,46-55

[3] M. Cochinwala, V. Kurien, G. Lalk, and D. Shasha, "Efficient Data Reconciliation", Information Sciences, 2001,137, 1-4

[4] M. Bilenko, R.J. Mooney, W.W. Cohen, P. Ravikumar, and S.E. Fienberg, "Adaptive Name Matching in Information Integration", IEEE Intelligent Systems, 2003,18(5), 16-23

[5] Ji Zhang, Tok Wang Ling, Robert. M. Bruckner, Han Liu, "PC-Filter: A Robust Filtering Technique for Duplicate Record Detection in Large Databases", Lecture Notes in Computer Science,2004,486-496

[6] Oktie Hassanzadeh, Fei Chiang, Hyun Chul Lee, Ren´ee J. Miller, "Framework for Evaluating Clustering Algorithms in Duplicate Detection", VLDB, 09

[7] Brendan J. Frey, Delbert Dueck, "Clustering by Passing Messages between Data Points ". Science, 2007,vol.315.973-976

[8] Huan Liu, Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge And Date Engineering,2005,17(4),491-502

[9] Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao, "A Variance Minimization Criterion to Feature Selection Using Laplacian Regularization", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001,33(10), 2013-2025

[10] Ahmed K. Farahat Ali Ghodsi Mohamed S. Kamel, "An Efficient Greedy Method for Unsupervised Feature Selection", IEEE International Conference on Data Mining, 2011, 161-170

[11] Qinbao Song, Jingjie Ni, Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions on Knowledge And Date Engineering, 2013,25(1).

[12] O. L. Mangasarian, W. H. Wolberg, "Cancer diagnosis via linear programming", SIAM News, 1990, 23(5),1-18.

[13] G. Demiroz, H. A. Govenir, and N. Ilter, "Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals", Aritificial Intelligence in Medicine,1998,13(3),147-165

[14] X. He, D. Cai, and P. Niyogi. “Laplacian score for feature selection”. In: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2005.

[15] Pabitra Mitra, C.A.Murthy, and Sankar K. Pal, “Unsupervised Feature Selection Using Fature Similarity”, IEEE Transactions on Knowledge And Date Engineering, 2002, 24(3),301-312

[16] Daoqiang Zhang, Songcan Chen and Zhi-Hua Zhou, “Constraint Score: A New Filter Method for Feature Selection with Pair wise Constraints”, Pattern Recognition,2008,41(5):1440-1451.

[17] C.M. Bishop. “Neural Networks for Pattern Recognition”. Oxford University Press, 1995.

[18] S. Kirkpatrick, C.D. Gelatt, Jr. and M.P.Vecchi, “Optimization by Simulated Annealing”, Science, 1983,220,671-680

Diao Xingchun, born in Taixing Jiangsu, received a master’s degree of computer science, from National University of Defense and Technology, Changsha, China. he is now a researcher in PLA University of Science and Technology, at Nanjing, China., and he is majored at data engineering.

Zhou Xing, born in Guangan Sichuan, received a master’s degree of computer science from PLA University of Science and Technology, at Nanjing China, he is now a phd candidate of computer science in PLA University of Science and Technology, majored at data engineering.

Cao Jianjun, born in Huicheng Shandong, received a doctor’s degree of computer science from Ordnance Engineering College, at Shijiazhuang China, he is now an engineering in PLA University of Science and Technology, at Nanjing, China., and he is majored at data engineering.